@maebert
#pycon2016
NERD TRIVIA

ROUND 1
Proposed withdrawal of Greece from the eurozone
GRINDERS

HACKERS WHO IMPLANT ELECTRONICS INTO THEIR BODIES
LATINX
GENDER-NEUTRAL FORM OF 
LATINA & LATINO
DAD BODS

SMALL FAT DEPOSITS ON OTHERWISE ATHLETICALLY BUILT MALES
Combination of Blanket and Scarf
I don’t know why. It’s a perfectly cromulent word.
PERFECTLY CROMULENT
HOW

DICTIONARIES

ARE MADE
SAYS URBAN DICTIONARY.COM:

TOP DEFINITION

**mansplain**

Stating verifiable facts that are inconvenient to the feminist worldview.

by Mansplainer April 17, 2014

[like 4876 dislike 2849]

BUY THE MUG

2

**Mansplain**

Telling a woman that she's wrong, even when she actually is.
“Her gown was of white satin worked with gold, and had long open pendent sleeves, while from her slender and marble neck hung a cordeliere—a species of necklace imitated from the cord worn by Franciscan friars, and formed of crimson silk twisted with threads of Venetian gold.”

— WH Ainsworth, Windsor castle
Quantitative Analysis of Culture Using Millions of Digitized Books

Jean-Baptiste Michel,1,2,3,4,5† Yuan Kui Shen,2,6,7 Aviva Presser Aiden,2,6,8 Adrian Veres,2,6,9 Matthew K. Gray,10 The Google Books Team,10 Joseph P. Pickett,11 Dale Hoiberg,12 Dan Clancy,10 Peter Norvig,10 Jon Orwant,10 Steven Pinker,5 Martin A. Nowak,4,13,14 Erez Lieberman Aiden1,2,6,14,15,16,17†

We constructed a corpus of digitized texts containing about 4% of all books ever printed. Analysis of this corpus enables us to investigate cultural trends quantitatively. We survey the vast terrain of ‘culturomics,’ focusing on linguistic and cultural phenomena that were reflected in the English language between 1800 and 2000. We show how this approach can provide insights about fields as diverse as lexicography, the evolution of grammar, collective memory, the adoption of technology, the pursuit of fame, censorship, and historical epidemiology. Culturomics extends the boundaries of rigorous quantitative inquiry to a wide array of new phenomena spanning the social sciences and the humanities.

pages of 1208 books. The corpus contains 386,434,758 words from 1861; thus, the frequency is $5.5 \times 10^{-5}$. The use of “slavery” peaked during the Civil War (early 1860s) and then again during the civil rights movement (1955–1968) (Fig. 1B).

In contrast, we compare the frequency of “the Great War” to the frequencies of “World War I” and “World War II”. References to “the Great War” peak between 1915 and 1941. But although its frequency drops thereafter, interest in the underlying events had not disappeared; instead, they are referred to as “World War I” (Fig. 1C).

These examples highlight two central factors that contribute to culturomic trends. Cultural change guides the concepts we discuss (such as “slavery”). Linguistic change, which, of course, has cultural roots, affects the words we use for those concepts (“the Great War” versus “World War I”). In this paper, we examine both linguistic changes, such as changes in the lexicon and grammar, and cul-
How many words are missing?

One Million Words
THE PLAN
WHAT COULD POSSIBLY GO WRONG?
MISSING WORDS

DETECT LANGUAGE

SAVE TO DB

PREPROCESSING

DETECT FRD

1.8M

“...” → FRD

FRD

12 YEARS

BING SEARCH

HTML PARSING

8 MONTHS

“...” → FRD

FRD

6 DAYS

html

15 HOURS

html

12

x50
NERD TRIVIA

ROUND 2
S3
EC2
ElasticSearch
EC2
LAMBDA
h'ors d'oeuvres

i collect bizarre animal confectionery dioramas

janky

kryogenkryokonitekryoscopy

list of unusual deaths

macÃƒÆ’Ã‚Â£Ãƒâ€”naaaaa

okcupid uranium-thorium dating

paenismus

quaestuary

revoltingly viviparous

shit my pt says

the dogs of slavery, misgovernment, and ostracism

outsnark

vésigniéîte

woggle

xenofeminism

yebo

zyxnoid
adeiladu  bialya  conquistar  DL 3 Aminoisobutyric

eu oi oa ou  frappul  galen's bondage etymologies

h’ors d’oeuvres  i collect bizarre animal confectionery dioramas

janky  kryogenkryokonitekryoscopy  list of unusual deaths

macÃƒÆ’Ã‚Â‚Ã£Ãƒâ€  naaaaa  okcupid uranium-thorium dating

paenismus  quaestuary  revoltingly viviparous  shit my pt says

the dogs of slavery, misgovernment, and ostracism  outsnark

végaigniéite  woggle  xenofeminism  yebo  zyxnoid
conquistar
DL 3 Aminoisobutyric
galen's bondage etymologies
i collect bizarre animal confectionery dioramas
janky
kryogen kryokonite kryoscopy
list of unusual deaths
macÃƒÆ’Ã‚Â£ÃƒÂ¢€
naaaaa
okcupid uranium-thorium dating
paenismus
quaestuary
revoltingly viviparous
shit my pt says
the dogs of slavery, misgovernment, and ostracism
outsnark
vésigniéite
woggle
xenofeminism
yebo
zyxnoid
I collect bizarre animal confectionery dioramas.
def valid_term(term):
    words = term.split()
    exclusion_rules = (  
        any(len(word) > 15 for word in words),
        len(words) > 5,
        any(c in term for c in "[]),?:1234567890"),
        sum(ord(c) > 255 for c in term) > 2,
        all(len(word) < 3 for word in words)
    )
    return not any(exclusion_rules)
from collections import defaultdict

def trigram_freq(text):
    trigrams = [text[k:k+3] for k in range(len(text)-2)]
    freq = defaultdict(float)
    for trigram in trigrams:
        freq[trigram] += 1.0 / len(trigrams)
    return freq
Detecting A Language

languages = {
    "english": trigram_freq("a quick brown fox jumps..."),
    "italian": trigram_freq("ma la volpe col suo balzo..."),
    "klingon": trigram_freq("SoH 'ej SenwI' rIlwI' je ...")
}

def detect_language(text):
    scores = defaultdict(float)
    for trigram, text_freq in trigram_freq(text).items():
        for lang, lang_freq in languages.items():
            scores[lang] += lang_freq[trigram] * text_freq
    return max(scores, key=scores.get)
Detecting a Language (Ghetto Remix)

stopwords = """all just being over both through its before herself had should to only under ours has do them his very they not now him nor did"""".split()

def is_english(text):
    words = text.split()
    n_stopwords = sum(w in stopwords for w in words)
    return float(n_stopwords) / len(words) >= 0.12
MACHINE LEARNING
Romney is trying to prevent a stampede to Trump of Vichy Republicans, collaborationists coming to terms with the occupation of their party.

vs.

There are a lot of elected Vichy Republicans who don't know how to do anything but lose, or kowtow to an authority figure.
WHAT PEOPLE THINK MACHINE LEARNING IS ABOUT

90% Alluring brain-inspired algorithms
10% "Big data"

WHAT MACHINE LEARNING IS ACTUALLY ABOUT

25% Picking the right features
70% Having clean, strong data
5% Choosing the right algorithm
TEXT CLASSIFICATION

TEXT

VECTORISER (.32, .14, .78)

TRAINING SET
(.32, .14, .78, 1)
(.2, .72, .03, 0)
(.44, .05, .91, 1)

(.15, .7, .65, ?)

CLASSIFICATION MODEL

.91
TEXT CLASSIFICATION IN PYTHON

training_data = [
    ("Horror vacui is a latin expression that means 'fear of emptiness'", 1),
    ("She put the cordeliere down next to a cap of black velvet faced with white satin", 0),
    ("Abusing anyone, I was told, violated Islamic tenets against zulm, or cruelty.", 1)
]

sentences, classes = zip(*training_data)
# Text Classification in Python

```python
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB

tfidf = TfidfVectorizer()
vectors = tfidf.fit_transform(sentences)
classifier = MultinomialNB()
classifier.fit(vectors, classes)

# Predict something
s = "A rose is a rose"
vectorised_s = tfidf.transform([s])
classifier.predict(vectorised_s)
```
PHAD SETTING UP AN AWS LAMBDA ENV ON EC2 FO
$ yum -y install blas lapack atlas-sse3-devel
$ virtualenv ~/stack; source ~/stack/bin/activate
$ dd if=/dev/zero of=/swapfile bs=1024 count=1500000
$ mkswap /swapfile; chmod 0600 /swapfile; swapon /swapfile
$ pip install numpy scipy pandas sklearn
$ strip `find ~/stack/lib/python2.7/ -name="*.so"`
$ pushd ~/stack/lib/python2.7/site-packages/
$ zip -r9q ~/lambda.zip * ; popd
$ aws s3 cp ~/lambda.zip s3://my_bucket/lambda.zip
$ aws lambda update-function-code --s3-bucket my_bucket \ --s3-key lambda.zip --function-name lambda_function
FULL AWS LAMBDA WALKTHROUGH:
BIT.LY/ML.AWS.LAMBDA

ALL THE CODE:

SERAPIS
SCALABLE WORD GOBBLER

GITHUB.COM/SUMMER.AI/SERAPIS